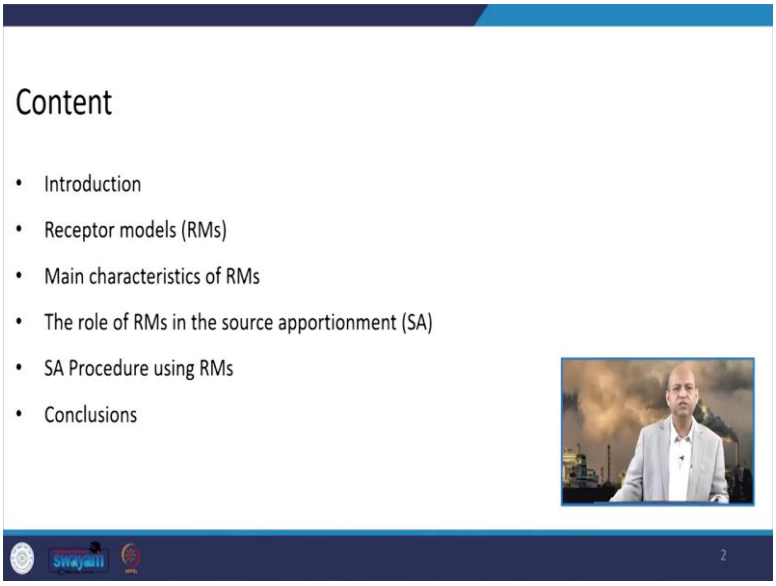
Air pollution and Control
Professor – Bhola Ram Gurjar
Department of Civil Engineering
Indian Institute of Technology – Roorkee
Lecture – 28
Source Apportionment using Receptor Modeling

Hello friends. Today we will discuss about source apportionment using receptor modeling. As you may recall, we have discussed general aspects of source apportionment, what is the source apportionment, what are different techniques or tools for doing the source apportionment. Basically, when we talk about a particular urban air shade or a particular location and we want to know that whatever concentration we are getting of a particular pollutant, then how much it is coming from a particular source.

Whether it is coming from let us say 2 wheelers or 4 wheelers or from power plant or from domestic sources, so to specify the source means the contribution of a specific source to the receptor means who is receiving the air concentrations of that particular pollutant. So, what is their absolute and relative contribution as per the source and to the receptor?

So, that relationship we want to know through the source apportionment studies. Because only after that we would be able to know that a particular pollutant is coming from a particular dominating source. So, that we can attack that source to control it whether through some policies, through some technologies or changing some fuel etc.

(Refer Slide Time: 1:48)

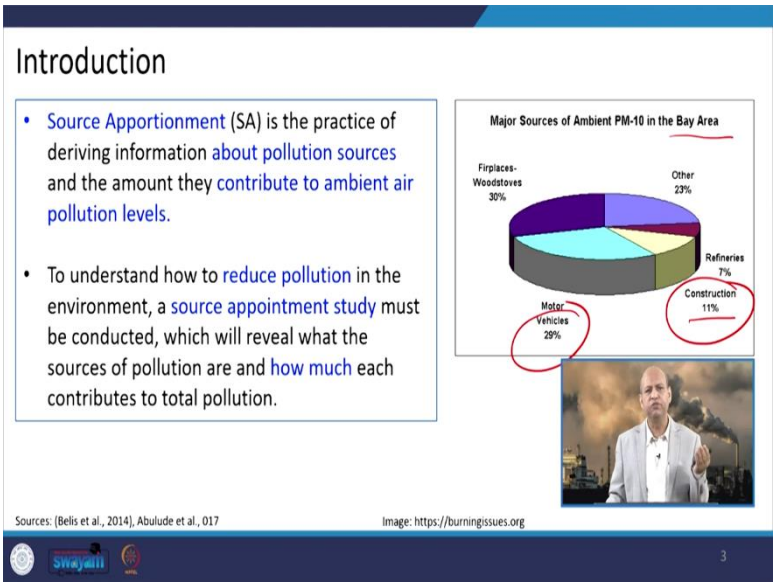


So, the content of today's lecture would be like, we will discuss little bit on introduction. Then what are the receptor models and the main characteristics of these receptor models are of

various kinds. So, we will discuss about them. Then the role of receptor models in the source apportionment. How do they play their role and the procedure of source apportionment using the receptor model.

So, the receptor model is the basic focus of today's lecture. To use them to do the source apportionment, otherwise there are several techniques but we will focus on receptor modeling and the conclusions thereafter we will summarize this lecture on the conclusions.

(Refer Slide Time: 2:25)

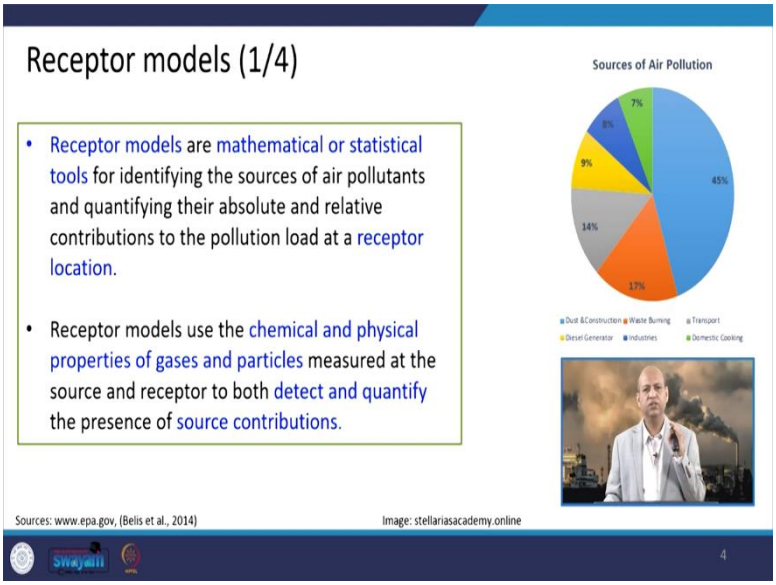


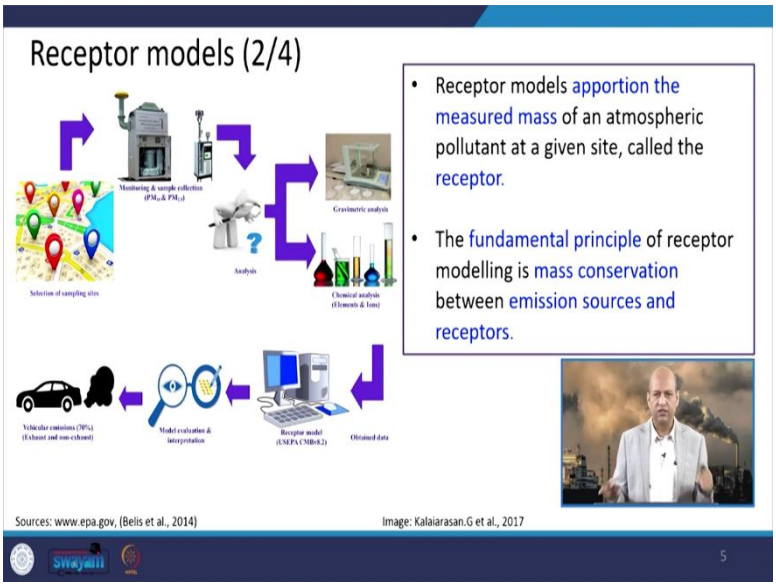
So, when we talk about source apportionment as you know, it is nothing but a particular exercise or practice of deriving information about pollution sources and the amount they contribute to ambient air pollution levels, means different sources they will contribute different amount of a particular pollutant. So, we will know how much amount is coming from a particular source. Here source ambient of PM_{10} in a Bay Area. So, 29 percent is from motor vehicle, 30 percent from fireplaces, wood stoves etc. Refineries contributing 7 percent, construction activity is 11 percent.

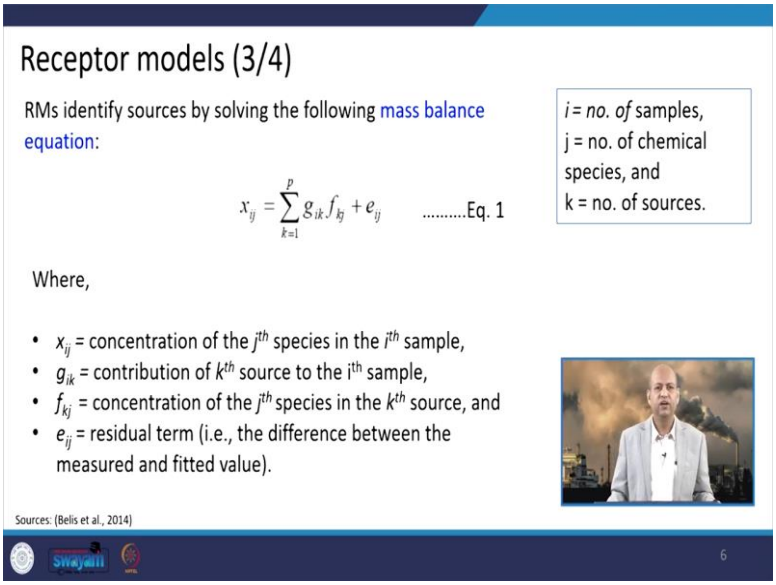
So, the PM_{10} is coming 29 percent from motor vehicles, 11 percent from construction. So, that is the source apportionment, means how much it is coming from different sources. So, to understand how to reduce the pollution in the environment is source apportionment study is a must activity, you can say. It needs to be conducted because it will reveal what the sources of the pollution are and how much each of the contributing source is contributing to the particular air pollution level.

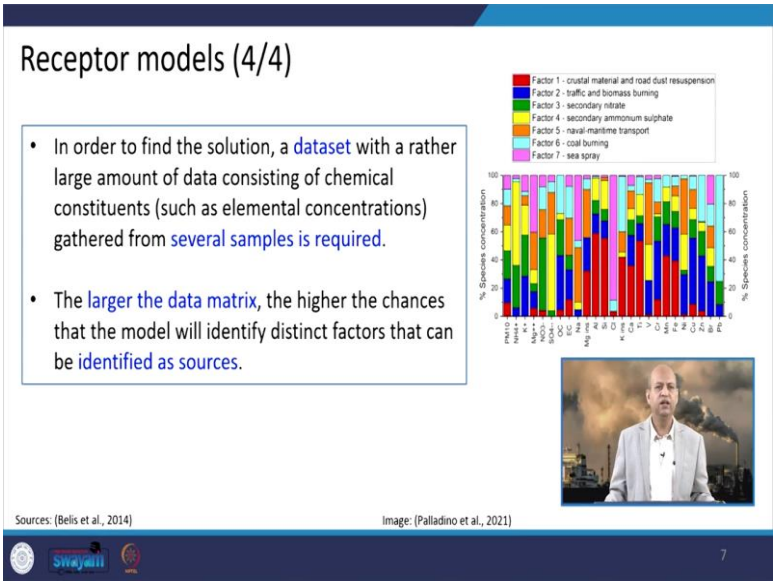
So, when we talk about receptor models because we are focusing in this lecture on receptor modeling otherwise there are several other techniques for source apportionment. So, when we talk about receptor models they are nothing but mathematical or statistical tools like equations and relationships to identify the sources of air pollutants and quantifying them in terms of absolute or relative contributions of the pollution load at a receptor location.

(Refer Slide Time: 4:10)









Wherever location we are interested to find the concentration and the source contribution and these receptor models basically use the chemical and physical properties of gases and particles because pollutants can be in gaseous form or particulate matter form and then they are measured at a source and receptor to both detect and quantify the presence of source contributions. For example, here you can see, this air pollution source like 45 percent is coming from dust and construction activities. Industries is contributing around 8 percent and then the waste burning is contributing around 17 percent.

So, that kind of distribution of relative contribution of different sources of a particular pollution can be there. Well, when we again look at the receptor models then basically these receptor models apportion the measured mass of an atmospheric pollutant at a given site or location that is known as the receptor, that is why the name is receptor modeling.

And the fundamental principle of the receptor modeling is a mass conservation between emission sources and the receptor because we assume that whatever the mass of that particular pollutant has been released from the source, it has come to the receptor. So, that kind of assumption we work upon.

And these are the equations which we use for receptor modeling where you know sources are identified and this equation is solved for the mass balance equation, we call it. So, here you can see X_{ij}, j equals summation of g_{ik} into f_{kj} plus e_{ij} and here this X_{ij} is nothing but the concentration of the j^{th} species in the i^{th} sample. So, there may be several samples and there may be in each sample different species like PM_{10}, SO_2, NO_2. So, when we talk about particular species, in a particular sample and we calculate the concentration of that or monitor the concentration of that particular species.

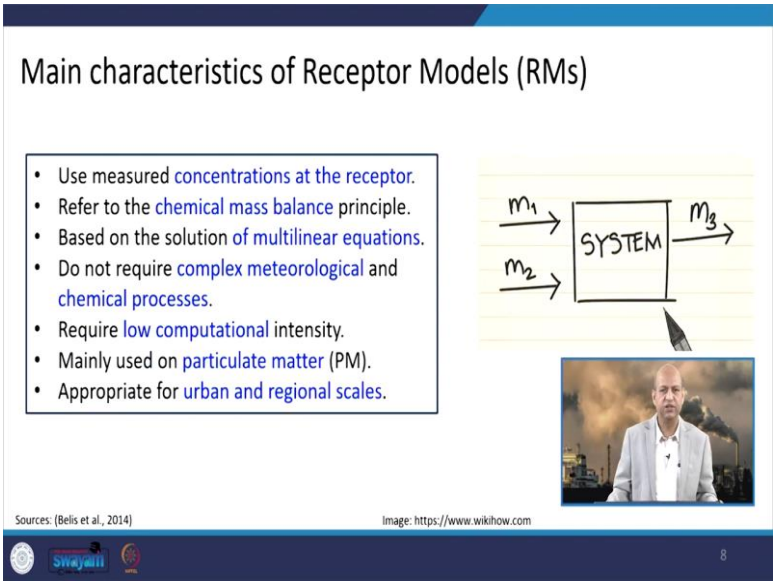
$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}$$

Then we will see that the contribution of k^{th} source because that will be coming from n number of sources, so the contribution of the k^{th} source in the j^{th} sample that particular sample and f denotes the concentration of j^{th} species in the k^{th} source. So, the multiplication of these all permutation combinations happen and we multiply and sum up that to calculate the total amount and then there may be some gap, because observations and calculations and the total concentration may not be exactly same.

So, then we also use this residual term e_{ij} so residual term that is the difference between the measured and the fitted value. Sometimes it is used, sometimes it is not used depending upon the situation. And, when we go further, then in order to find the solution a data set with a rather large number of data consisting of chemical constituents such as elemental concentrations and gathered from several samples it is required.

Large number of data is available then good robust evaluation happens. So, the large the data matrix, the higher the chances that the model will identify distinct factors that can be identified as the sources. So, identification of sources becomes easier in that case when we have lot of data.

(Refer Slide Time: 7:43)

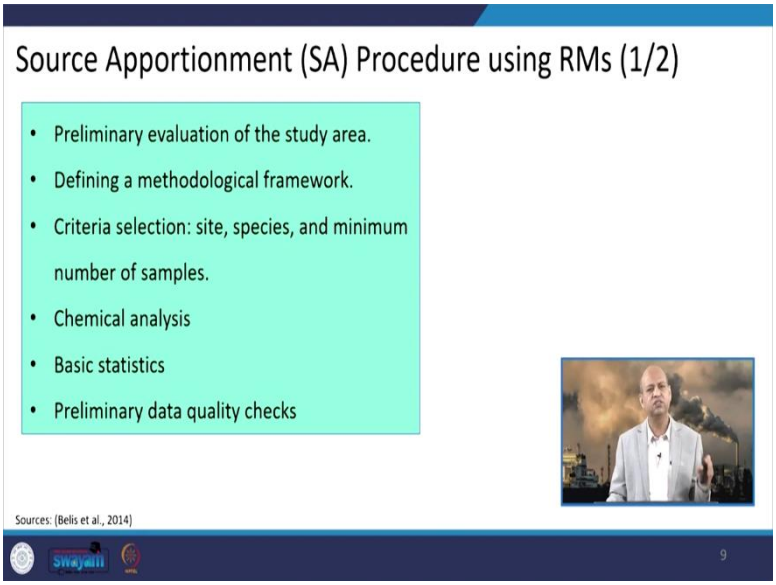


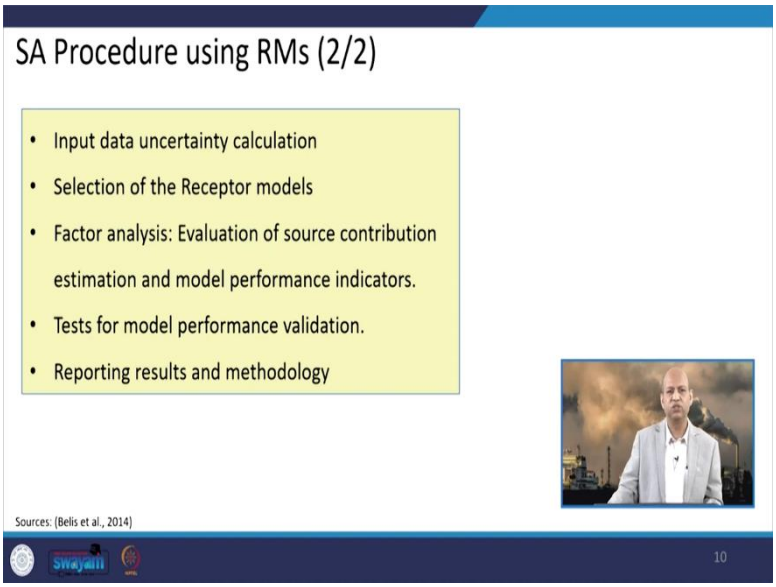
And when we look at the main characteristics of receptor models then there are number of characteristics like the use measured concentrations at the receptors. So, these receptor models they use the concentrations which are measured at the receptor or location, receptor location. And it refers to the chemical mass balance principle basically.

Whatever mass balance is there input and output that should be the same. Based on the solution of multi linear equations, so different variables are there and it does not require like complex meteorological and chemical processes which we need in some other models, those complex models.

So, this is easier one and it requires low computational intensity. So, that is why it is preferred one and mainly it is used on particulate matter because that mass balance thing is properly suitable for this. Otherwise for gaseous species some reactions may happen which we generally do not account in this particular technique. The appropriate for urban and regional scales for that, in that sense it is good results are received by receptor modelling, on those scales.

(Refer Slide Time: 8:57)





And when we see, what is the procedure? Source apportionment, the procedure using receptor modeling so again different steps are there. For example, we do first of all preliminary evaluation of the study area. So we will look at the study area and the different features aspect of that study area, we note down and discuss about that. Then we define a methodological framework depending upon what kind of study area it is.

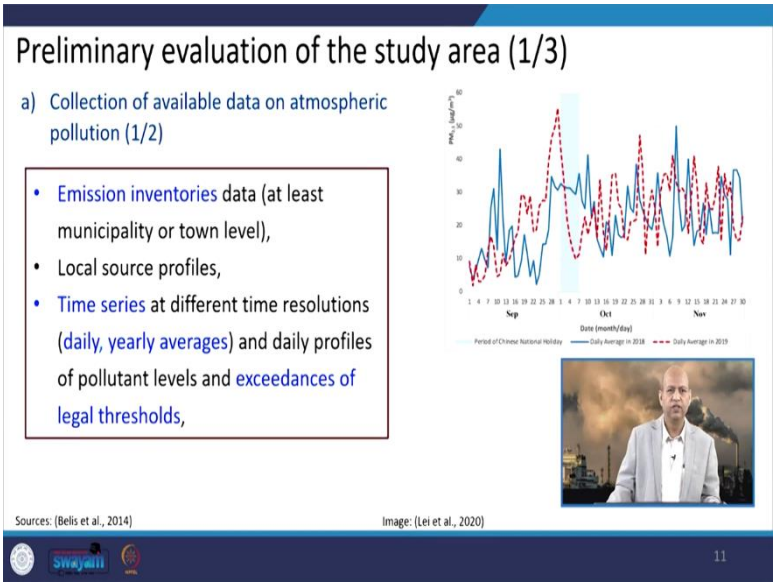
It is coastal area or it is land logged area hilly area. Accordingly, specific requirements may be there, so the framework is designed based on that preliminary evaluation of the study area. Then the criteria selection like site, how to select the criteria? So, site or those criteria which are needed for these computational activities. So, site must be there, species or minimum

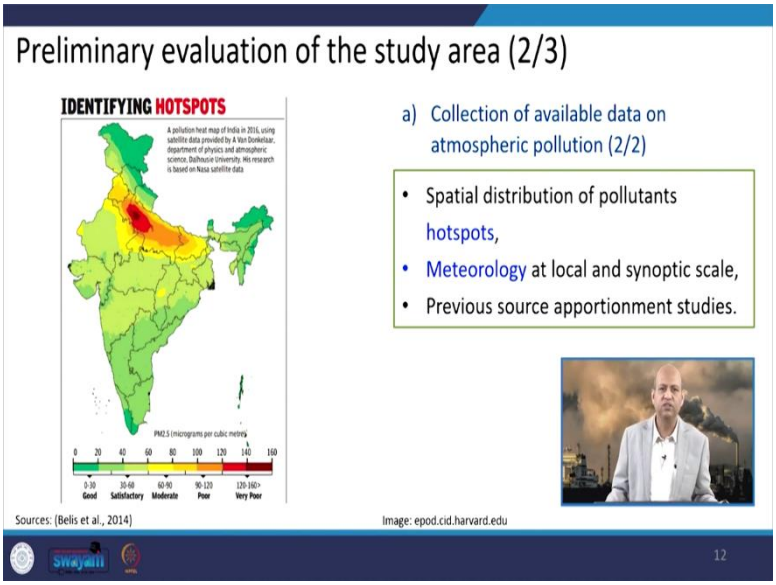
number of samples so these are the basic criteria which are selected for carrying out this source apportionment through using receptor modeling.

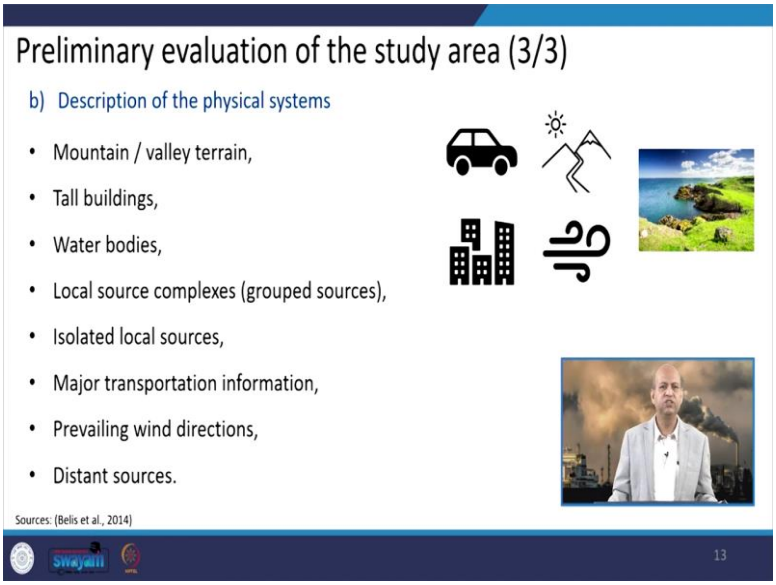
And the chemical analysis is very important because only then we will be able to know how much heavy metal is there, how much a particular species is there in the particulate matter. And then basic statistics because the relationship it uses for sensitivity analysis for uncertainty analysis, all those things basic statistics really help us and then the preliminary data quality checks are also needed.

Then the input data uncertainty calculations are to be done and selection of the receptor models is to be carried out. Then the factor analysis evaluation of the source contribution by factor analysis when we estimate and do model performance indicators and test of the model performance validation is also to be conducted or carried out. Then the reporting results and methodology is the last thing based on the those calculations and evaluations we do.

(Refer Slide Time: 10:59)







So, when we come to the first step like preliminary evaluation of the study area, important aspects are like emission inventories. At least like municipality area or town level area if some very basic emission inventories available then it will give us some insight that which sources are dominating. Although it is not necessary that at the receptor only those dominating sources will be there because it will change depending upon the location, it will change.

It is not necessary, so, that is why a mission inventory based source apportionment may be different than this receptor modeling based one. So, local source profiles will see, what are the different sources, what are their profile, intensity, which kind of fuel they are using all those things.

Then the time series at different time resolutions like daily variation or yearly averages, those kind of averages values we need to compute and the profiles based on those pollutant levels

whether it is daily or annual level, we compare with those threshold values, you can say like (NAAQS), National Ambient Quality Standards whether they are exceeding those standards or not or lower than those. Those kind of information help us to carry out which space is more important to study.

Then we look at like collection of available data on atmospheric pollution. We continue in terms of special distribution of the pollutants, which are the hot spots where lot of air pollution is being emitted. Then meteorology at the local and synoptic scale, so you can see means what are their wind velocity or those adiabatic lapse rate related comparison with the environmental lapse rate and those stability classifications, moisture, temperature all those things we need to have.

Then the previous source apportionment studies if they are available, then we have to see because land use, land planning changes over the period of time. So, the receptor also changes in terms of getting different kind of species of air pollutants. When we describe about the physical systems, then we see, of that steady area. Physical system means what are the topographical features are there.

There are mountains or valley terrain or like tall buildings are there, some street canyon effect is happening or not some water bodies are there or not because they will influence the meteorological parameters. Large water body is there then temperature gradient may be different and moisture content may be different.

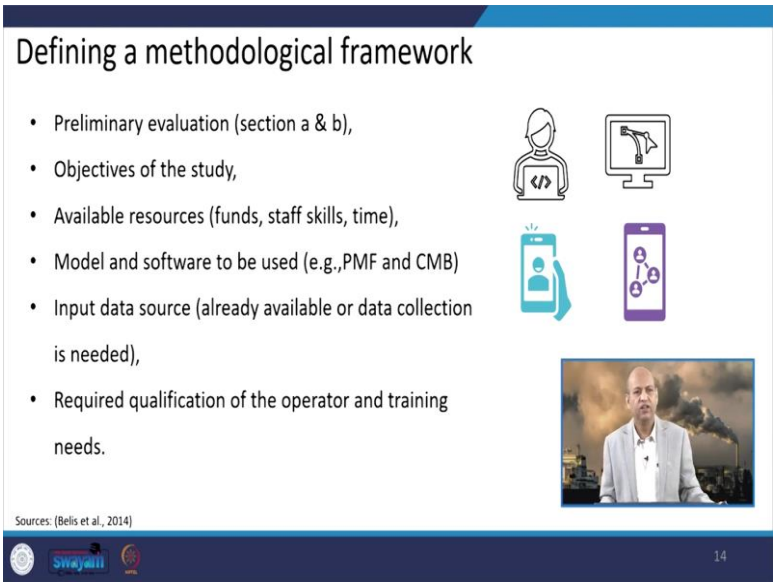
Then the local sources, which are very complex like different kind of sources are clubbed together. Then isolated local sources may be there. So, some combined sources may be available some isolated sources may also be there like distinct sources. Suppose power plant is there so it is quite distinct source of pollution.

We know which fuel is being burnt in the power plant like coal then we easily know that what kind of pollutants are coming from that. But there are mix like in industrial area several kind of fuels are being burnt then every kind of vehicle is coming from diesel based or LPG based or CNG based or gasoline based. So, it is a mixed one, that kind of thing may happen.

Then the major transportation information, all categorization of vehicle category there, fuel type, technology type, 2-stroke, 4-stroke all those things. Then the prevailing wind direction is important because it will give us an idea that may be some additional source is coming from

the up wind direction. So, the wind direction is very important and the distinctive sources through back trajectory we can calculate, if we know these wind related profiles.

(Refer Slide Time: 14:58)



Then when we try to define these methodological framework so the preliminary evaluation which we do accordingly this a and b we have already discussed, then the objective of the study are noted down properly. So, that we do not deviate from the aims and objectives of the study.

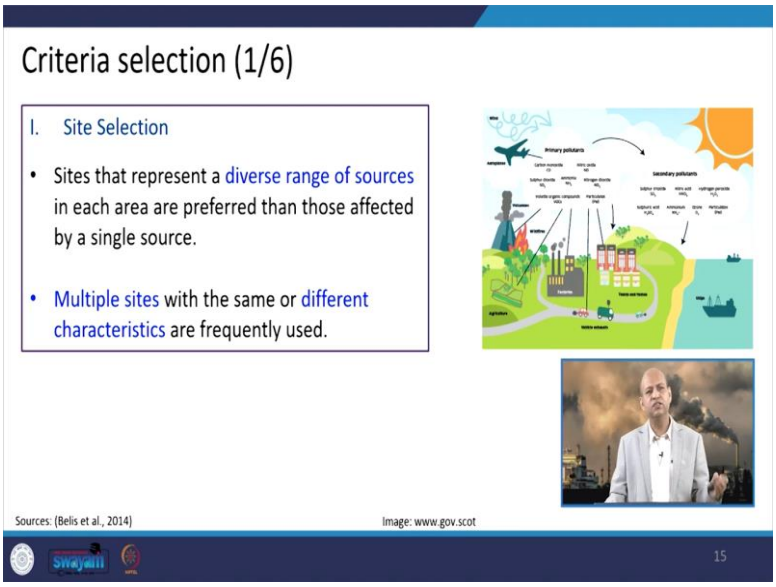
Then we go for available resources like how much funds is available? How much staff is there? And whether staff is highly skilled or low skill kind of staff? How much time is available to do? Because they will determine how much extent of this exercise we have to do. If you do not have large funds then we have to go for simple exercise of source apportionment. We cannot go for detailed in one.

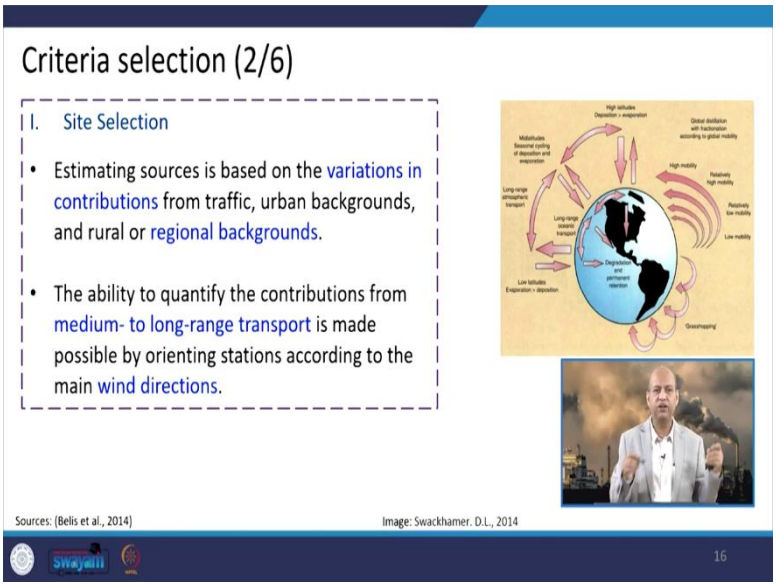
Then when we talk about like model and software which need to be used for this exercise of doing source apportionment through receptor modeling. Like (PMF), Positive Matrix Factorization to be done or not, chemical mass balance is to be done. So, which software, which model we are going to use. Then input data sources like which may be already available or maybe we have to collect those required data from different sources or by preliminary primary survey.

Then required qualification of the operator and training needs according to, because if you are using a software then the person should know how to deal with the software. You cannot just employ someone who does not know the software and then you just ask him to you know learn

and do. Then it will require lot of time, it will be wasted so many months and maybe the study is delayed.

(Refer Slide Time: 16:33)

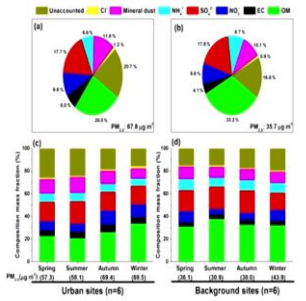




Criteria selection (3/6)

II. Species selection (1/2)

- The chemical species to include in the analysis should be selected according to the study objectives.
- Traditional receptor models focus on a small number of chemical species to represent the majority of particulate mass.



Sources: (Belis et al., 2014)

Image: Liu. Z. et al., 2017

Criteria selection (4/6)

II. Species selection (2/2)

- The major ions (sulphate, nitrate, and ammonium) and the carbonaceous fraction (TOC, OC, and EC), as well as a few other elements whose absolute and relative concentrations or specific ratios are used to identify sources.
- NO^{3-} is related with particles larger than those associated with SO_4^{2-} and ultra-fine particles (UFP) with gasoline and diesel exhausts but not with vegetation burning.

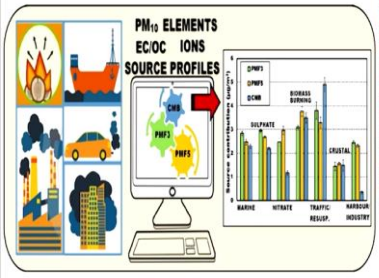
Sources: (Belis et al., 2014)

Image: https://nwfuel.ca, https://indiaclimatedialogue.net

Criteria selection (5/6)

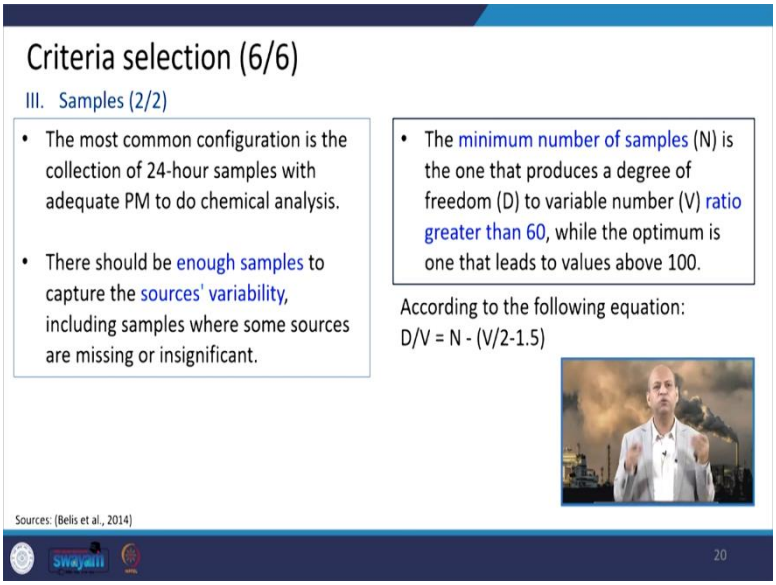
III. Samples (1/2)

- In practice, many samples are required to obtain results that are representative of the variety of conditions in the study area, including the variability of sources over time.
- Multivariate techniques only work properly with large numbers of samples as input.



Sources: (Belis et al., 2014)

Image: Cesari.D., 2016



Then the criteria selection is also like site selection, as we have discussed. So, sites that represent a diverse range of the sources in each area are preferred than those affected by a single source because diverse sources then these models really work well for different sources and the multiple sites with the same or different characteristics are frequently used. So, that is also very important criteria of the site selection.

Then when we go for estimating the sources which are based on variations in contributions from different sources like traffic, urban background, rural or regional background because then different pollutants will come according to the background, rural emissions are different than the urban. In rural, biogenic emissions may be more. The ability to quantify the contributions from medium to long range transport is also made possible by orienting stations according to the main wind direction because as you know, from wind direction some pollutant may come, as we have already discussed. So, that orientation is also very important.

Then when we talk about species selection which species are more important. So, the chemical species which are to be included in analysis that should be selected according to the study objective, otherwise we will waste lot of time if we cater all kind of pollutants and species. We should focus only those important species which are according to our objectives.

When we are going for acid rain related then better you go for SO_2 NO_x, you do not need to go for other pollutants. Traditional receptor models focus on a small number of chemical species to represent the majority of particulate masses. Because that gives lot of information about different characteristics of pollution sources.

Well then within this criteria selection, we talk about continuously about species selection. So, other important aspects are also there, like major ions like sulphate, nitrate, ammonium and the carbonaceous fraction like total organic carbon or elemental carbon, organic carbon as well as few other elements whose absolute and relative concentrations or specific ratios are used to identify sources that should be kept in mind.

Nitrate is related with these particles or particulate matter larger than those associated with sulphate, their contribution and ultra-fine particles with the gasoline and diesel exhaust but not with the vegetation burning. So, according to the source and objectives, we have to change the species selection criteria. Then the samples, how many samples we need to do. So, in practice many samples are required to obtain the results that are representative of the variety of conditions in the study area. Like it can include the variability of sources over time. Then multivariate techniques only work properly with large number of samples as input criteria.

So, that is also one important aspect depending upon how many samples we are collecting and for how many species and those kind of things are to be kept in mind. And, continuously when we talk about like samples so what happens the most common configuration is the collection of 24 hour samples with adequate particulate matter to do chemical analysis because you must have adequate quantity otherwise the chemical analysis will not be proper and there should be enough number of samples to capture the sources variability.

Otherwise some bias will happen, if you do not have adequate number of samples, large number of samples then some samples may miss, some important species or their pollution loading. And these samples which some sources are missing or insignificant they will be captured when large number of samples are there and from mathematical aspect, the minimum number of samples is the one that produces a degree of freedom to variable number V ratio greater than 60, D divided by V that is should be greater than 60. And while the optimum is one that leads to the values over above the 100. So, this is the equation which is used for calculation this particular ratio.

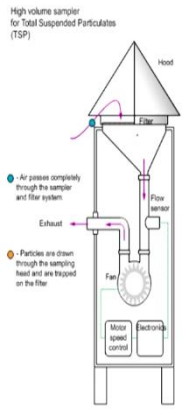
(Refer Slide Time: 20:51)

Chemical analysis (1/4)

I. Sampling systems (1/2)

- Chemical analysis of PM is commonly performed using filter-based methods.
- Filters are chosen based on the following criteria:

❖ Compatibility with analytical procedures, no interactions with samples, low impurity levels, and high efficiency.

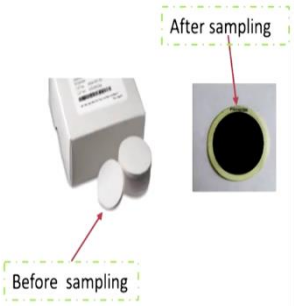


Sources: (Belis et al., 2014) Image: www.qld.gov.au

Chemical analysis (2/4)

II. Sampling systems (2/2)

- Commonly used filter matrices are pure quartz, coated quartz and teflon, nylon, polycarbonate, glass fibre and cellulose esters.
- Because carbonaceous aerosols must be analyzed at elevated temperatures, only quartz fibre filters are suitable for determining ions, elements, and carbonaceous fractions (organic and elemental) on the same sampling platform.



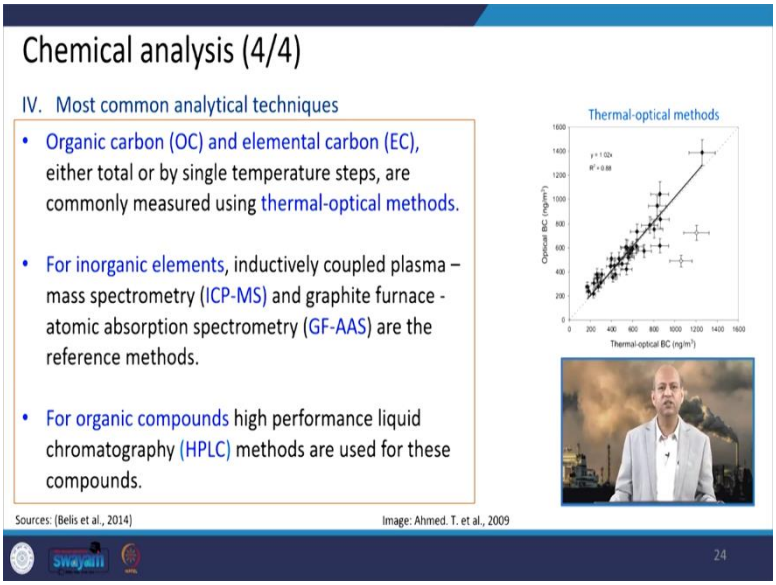
Sources: (Belis et al., 2014) Image: www.sigmaaldrich.com, Chakraborty and Mondal (2017)

Chemical analysis (3/4)

III. Local source profiles (fingerprints)

- Considering the resource requirement for local source characterization, fingerprints from previous work in similar domains or obtained from the source profile repositories (e.g., SPECIATE, US-EPA).
- Source-oriented monitoring stations can be used to characterize the source emissions if periods in which other sources influence the sample are excluded from the analysis.

Sources: (Belis et al., 2014) Image: www.envea.global



Now, when we do the chemical analysis, for that purpose first of all, we have to collect the sample. So, filter paper is used and this high volume sampler is the way the instrument which is simple instrument like air goes through this filter paper then the particulate matter is deposited on the surface of the filter paper, it is analyzed and we see what kind of different species are there.

So, the chemical analysis of particulate matter is commonly performed using the filter based method which is shown in this particular schematic diagram and filters are chosen based on the important criteria like compatibility with the analytical procedure. Whatever procedure we are using, so it should be compatible. It should not disturb the sampling procedure or species concentration or amount.

And it should not interact with the sample otherwise the values will change. It should be inert in that sense and it should not have impurity. If it has very very low impurity levels must be there otherwise our sampling, our analysis will be completely erroneous and it should also have high efficiency. So, the commonly used these filter matrices are pure quartz or coated quartz or Teflon, those kind of, nylon different kind of filters are available and because these carbonaceous aerosols must be analyzed at elevated temperature.

So, only quartz fibre filters we use for this particular determining different ions and elements and carbonaceous fractions in a particulate matter collection. So, this kind of filter paper before we do sampling after sampling we get because the particulate matter get deposited on the filter paper. So, this kind of ah difference is there. When we consider about local sources profile or fingerprints of the local sources, that is also very important aspect and information.

So, when we consider the resource requirement for local sources characterization or fingerprints of the local sources or sites from previous work, it is like similar domains or obtained from the source profile repositories which can be taken from different reports attributes are like United States EPA, Environmental Protection Agency. In case of India, you can go for Central Pollution Control Boards different reports etc. They can give us the background information about that particular local sources.

And the source oriented monitoring stations can be used to characterize the source emissions if periods in which other sources influence the sample are excluded from the analysis. So, that kind of cautiousness, we should take care. When we talk about like different analytical techniques then like organic carbon and elemental carbon, either total or by single temperature steps are commonly measured using the thermal optical methods.

And when we have to analyze these inorganic elements. Then ICP-MS, this Inductively Coupled Plasma Mass Spectrometry (ICP-MS), this is very much used and Graphite Furnace Atomic Absorption Spectrometry (GF-AAS), this is also used for this particular purpose of inorganic elements. Then if we go for analysis of organic compounds then High Performance Liquid Chromatography (HPLC) methods are used for these kind of compounds.

So, depending upon the different species, different chemical analytical methods or techniques we need to use, otherwise we will not be able to get the right results.

(Refer Slide Time: 24:26)

Basic statistics (1/8)

- It is good practice to make some summary plots and run some simple tests to gain an overview of the relationships between variables and how they change from sample to sample.
- Many commercial and free software applications are available that can carry out routine statistic tests (e.g., Statistica and MATLAB).

Sources: (Belis et al., 2014)

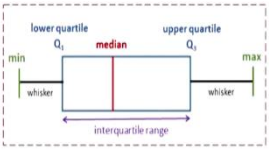
Image: www.pngegg.com

25

Basic statistics (2/8)

i. Central and dispersion statistics

- Box and whisker plots are useful to visualize central values of your variables (mean, median) and the dispersion of your data around the central values (quartiles, minimum and maximum values).



- These plots give a visual overview of the data spread that enables a preliminary assessment of the distribution.
- Visual tests of normality such as histograms, probability plots and normal probability plots are also useful.

Sources: (Belis et al., 2014)

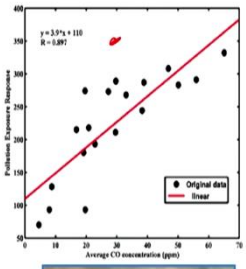
Image:

26

Basic statistics (3/8)

ii. Correlation matrices

- The correlation between variables can be visually assessed using scatter plots.
- This is particularly important for identifying abnormal data points that may impact the correlation.
- If many variables are involved, the Pearson correlation coefficient (R) and related statistics for every possible pair of variables is a useful exploratory technique, provided the influence of outliers has been evaluated.



Sources: (Belis et al., 2014)

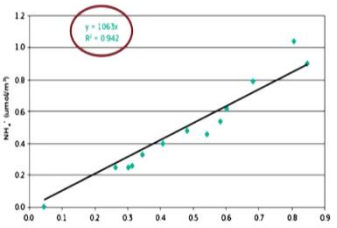
Image: Aliyu and Botai (2018)

27

Basic statistics (4/8)

iii. Linear regression

- Ordinary Least Squares regression is the simplest and quickest technique to more in-depth exploration of the association between two variables.
- Slope and the determination coefficient (R^2) provides useful preliminary information to describe the (linear) relationship between the variables considered.
- If R^2 is 0.9, the input variables explain 90% of the variation in the output variable.

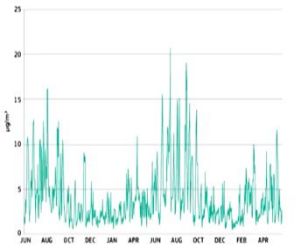


Sources: (Belis et al., 2014)

Basic statistics (5/8)

iv. Time trends

- Plotting time trends of the variables makes it possible to identify regular patterns in data.
- Seasonality, influence of the day of the week or extraordinary events that probably indicate the impact of specific sources influencing the study area for short periods like wildfires.
- In addition, when hourly data are available, characteristic daily profiles of certain species can be used to identify specific sources (e.g. a peak of traffic markers during rush hours).

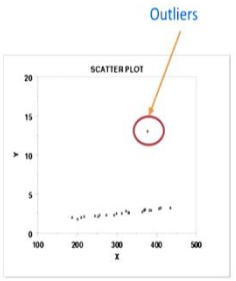


Sources: (Belis et al., 2014)

Basic statistics (6/8)

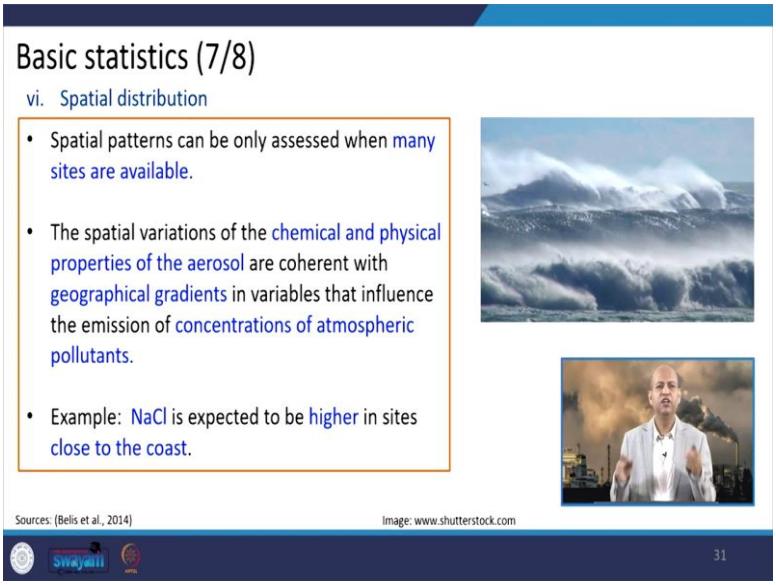
v. Outliers

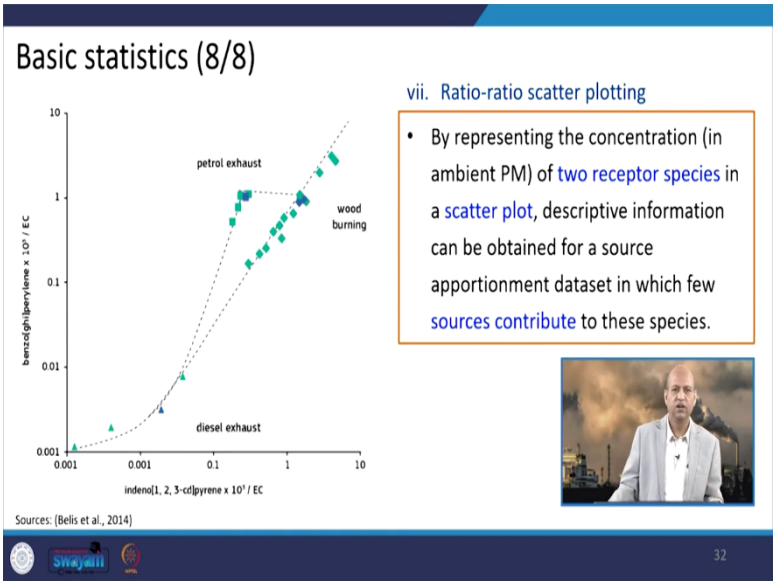
- Values that do not follow the distribution of data with similar characteristics are referred to as outliers.
- They may reflect genuine properties of the studied system or derive from measurement errors that are not relevant for the model.
- Grubb's test helps the analyst to decide whether data provide useful information on sources or whether they only introduce noise into the model.



Sources: (Belis et al., 2014)

Image: https://thejosevilson.com





Then basic static statistics are important for carrying out this study, so it is good to practice some summary plots and some simple tests to know the reliability or its robustness and the overview of relationships between variables and their changes with the sample to sample, if their variation because of some meteorological changes or changes in the way of taking the sample, all those changes are figured out in when we take different samples and analyze them through different statistical methods.

So, many commercial and free software are available for carrying out those statistical analysis and like Statistica, MATLAB, R kind of software. There are many kind of software you can use for that purpose and box and whisker plots are useful for visualizing central values of these variables like mean, median and the dispersion of data around the central values also like minimum, maximum values, all those things are to be carried out by those plots.

And these plots basically give a visual overview for the data spread that enables us for a preliminary assessment of the distribution because those plots are very good visual inspection happens through those plots. How much variation is there, how much contribution is there and this visual test of normality such as histograms or probability plots and normal probability plots they are very useful.

So, they should be carried out by using these basic statistical analysis. Then when we talk about different correlation matrices, so there are also, you can see these scatter plots may be there. So, they also give some information because identifying like abnormal data point suppose some point is here, so that is like outlier. We will see what is outlier, how it is identified?

So, variation of different values and abnormal data points, there can be done by these kind of correlation analysis and the variables which are involved like pearson correlation coefficient 'r' and related statistics for every possible pair of variable, it is useful for exploratory technique. It is a useful explanatory technique provided the influence of outliers has been evaluated and taken out.

Then otherwise, low correlation when 'r' value is less and nearer to 0 like 0.2, 0.1 and high correlation means 0.9, 0.95 then it is high correlation up to 1. So, those kind of values can give us those ideas about its variation and nearer to the right value. Then when we see least square regression, this is also important statistical parameter which gives the simplest and quickest technique to more in depth exploration of the association between two variables whether they are properly related or not.

And the slope and determination coefficient like r-square these provides useful preliminary information to describe linear relationships between variables considered whether they are linearly variable or not or they are completely unrelated. So, the r-square if it is 0.9, the input variables explain 90 percent of the variation in the output variable. So, those kind of you can see like here this nitrate and sulphate and this is ammonia whether they are related to each other or not.

So, those values are there, r-square is around 0.945. So, their relationship is very good in that sense, variation is very good. They are closely related. Then we talk about time series analysis, time trends that is also very important. So, the plotting time trends of the variables it makes possible to identify regular pattern in data, whether it is decreasing or increasing, having some peaks those kind of things.

So, if there is a peak then why that peak is occurring whether some local source has been in that particular time period or not. So, those kind of things, we can really see through these plots of the time trends. Then when we want to know the outliers, so these are the values which are not in the similar variation. So, they may reflect like genuine properties of studies, system derived from measurement errors that are not relevant to the model.

So, those outliers may be because of certain mistakes. So, by these basic analysis you can find out their pattern and those outliers can be removed otherwise they may give us wrong influence or wrong conclusions. When we talk about a spatial distribution, like geographical. So, special patterns can be only assessed by many sites available. Then variations are important and easy to detect.

The special variations of the chemical and physical properties of the aerosols are coherent with the geographical gradients in variables that influence the emissions of concentrations of atmospheric pollutants like NaCL. NaCL is expected to be higher in the sites close to the coast. So, that way these special variations or sites, properties can also give us idea which kind of pollutant may be higher in those locations.

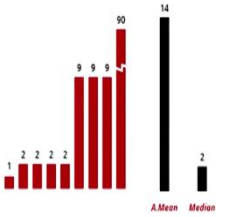
If we want to know these representative concentrations, how their two receptor species are there, so the ratio scatter for plotting can help to determine their contribution at a particular location.

(Refer Slide Time: 30:12)

Preliminary data quality checks (1/6)

i. Missing values

- Zero and negative values have no physical meaning, they should be removed from the input file or substituted with more appropriate values.
- The simplest option is to remove the row (sample) or column (species) from the input matrix. But this could result in the loss of important information.
- An alternative approach is to fill in missing values with estimated values, such as the mean, median, or geometric mean of the species' recorded concentrations in all samples from that study site.



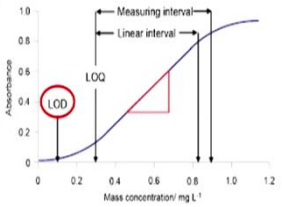
Sources: (Belis et al., 2014)

Image: https://jlmc.medium.com

Preliminary data quality checks (2/6)

ii. Values below the detection Limit

- Limit of detection (LOD) is a term used in analytical chemistry to describe the lowest concentration of the sample at which detection is possible.
- Values below the detection limit of the analytical method should be used if they are available.
- It has been suggested that only species with more than 50% of values below the detection limit should be included in the data set, as this is adequate to provide information about sources.



Sources: (Belis et al., 2014)

Image: https://www.efim.eu

Preliminary data quality checks (3/6)

iii. Signal-to-noise

- The power ratio between a desired signal (S, meaningful information) and background noise (N, unwanted signal) is known as the signal-to-noise ratio (S/N).
- The signal-to-noise ratio is useful for classifying variables according to the information they supply for the source identification analysis.

❖ Variables are to be excluded from the analysis

If, $S/N < 0.2$

❖ Variables are to be suitable for the analysis

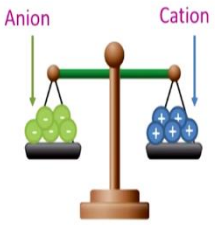
If, $0.2 \leq S/N < 2$

Sources: (Belis et al., 2014)

Preliminary data quality checks (4/6)

iv. Mass closure and ion balance

- To match the masses or the electric charges of species.
- In the first case, mass closure is accomplished by comparing the mass of particulate matter (PM) to the sum of the masses of the major chemical components.
- It is feasible to evaluate deviation from neutrality by comparing the sum of anion equivalents to the sum of cation equivalents and showing values in a graph assist in the identification of samples with an abnormal ionic composition.



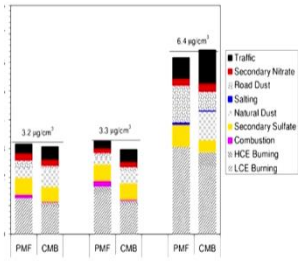
Sources: (Belis et al., 2014)

Image: https://smart-fertilizer.com

Preliminary data quality checks (5/6)

v. Analysis of consistency in time and space (1/2)

- Changes in sampling methodologies or analytical techniques may create disruption in time series that must be duly considered during data elaboration.
- Comparing time series from multiple locations assists in the identification of unusual patterns.



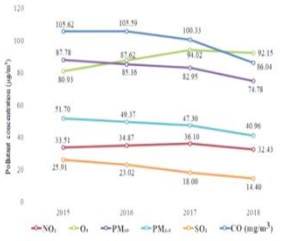
Sources: (Belis et al., 2014)

Image: Chen et al., 2011

Preliminary data quality checks (6/6)

v. Analysis of consistency in time and space (2/2)

- In order to populate a dataset with an appropriate number of samples it may be necessary to collect data for more than one year.
- However, species and other variables collected during different years may show different relationships.
- In order to check these patterns before running the analysis, scatter plots or time trend plots are useful.



Sources: (Belis et al., 2014)

Image: Liu, Y., Zhou, Y. & Lu, J., 2020

Then when we talk about preliminary data quality checks, so missing values or 0 and negative values which have kind of no meaning, so they can be removed by these kind of analysis. So, the simplest option is to remove those raw samples which are having these kind of values and as an alternative approach to fill these missing values with estimated values such as the mean, median or geometric mean of the species because 0 may be there when instrument is not working.

So, that means that value 0 is not right but because of certain reason that 0 is coming. So, we have to remove them and then the missing value has to be filled by those mean values because we can know through the trend or how much value can be there. So, those missing values can be filled by our expert opinion with the help of those data.

Then when we talk about which values are good and detection limits are or below detection limit whether they are important or not, so the limit of detection is a term which is used in analytical chemistry to describe the lowest concentration of the sample at which destruction is possible. Below that it will not be possible, so the values below the detection limit of the analytical method should be used if they are available.

Otherwise it is not possible. So, it has been suggested that only species with more than 50 percent of value below the detection limit should be included in the data set. As this is adequate to provide information about the sources. So, that detection limit and those kind of information and concept is very important to learn.

Then signal to noise, some information may be like noisy means that is again kind of outliers or they are not really important but they are coming from different sources. So, the power ratio between desired signal meaningful information that is desired signal S and the background noise and unwanted signal which has come because of some other sources which is not the regular source that is just outlier kind of thing.

It is known as the signal to noise ratio and this signal to noise ratio is useful for classifying variables according to the information they supply for the source identification analysis. So, variables which need to be excluded when this S / N is less than 0.2 that means N is very high in that number and the variables which are suitable for inclusion for the analysis are those for which S / N is greater than 0.2 but less than 2.

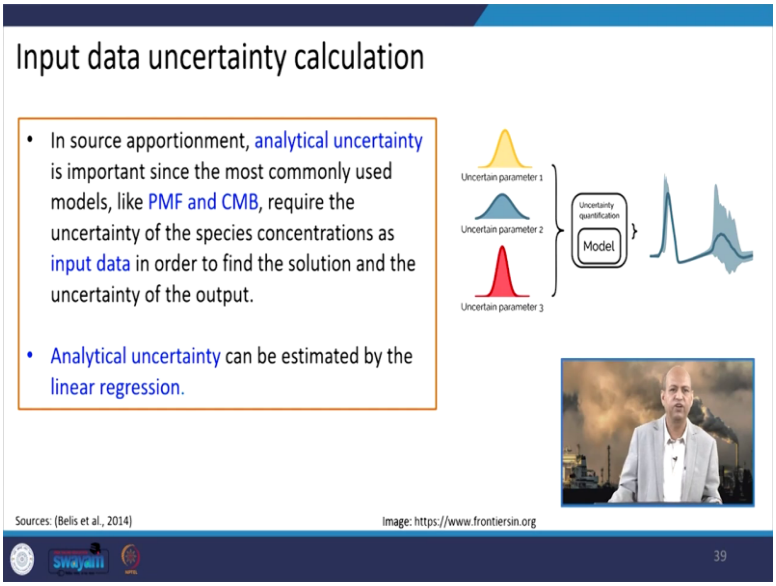
So, those kind of values we should know. Then when we talk about mass closure or ion balance. So, this again you have to see like there feasibility or evaluation we do depending upon

equivalence of anions, cations. So, their summation should be equal otherwise we have to fill those errors by calculations related to like modeling efforts.

When we do analysis of consistency in time and space so the sampling methodology changes, our analytical techniques may also create some disruption in time series. So, that must be duly considered during data elaboration otherwise again our interpretation may be wrong. So, the comparing time series from multiple location assist in identification of unusual patterns.

So, if we have multiple locations monitoring sites, if one site is not giving good information then nearer site may be taken as the base case so that it can you know fill up those mistaken kind of information. When we do like time series analysis, so again you can have the species and other variables collected during different years and you can have their variation, they are going up or down, what kind of regions are there for different patterns. So, those patterns give us some important information about their variation depending upon those factors.

(Refer Slide Time: 34:12)



When we talk about source apportionment like uncertainty calculation sensitivity analysis. So, those are also important thing and by this positive metric factorization or chemical mass balance, they are required for uncertainty of the species concentrations as input data in order to solution of uncertainty of the output. So, those values is very important. And analytical uncertainty can be estimated by the linear regression. So, that is one important way of doing that.

(Refer Slide Time: 34:46)

Selection of the Receptor Models (1/3)

Type	Examples
Exploratory methods	Enrichment factor, tracer method, incremental approach
Chemical Mass Balance	EPA CMB 8.2
Eigenvector-based models	PCA, UNMIX
Factor analysis without constrainits	FA, APCFA
Positive matrix factorization	PMF2, EPA PMF v3
Hybrid trajectory-based models	CPF, PSCF
Hybrid expanded models	PMF solved with ME-2, COPREM

Various types of Receptor models

CMB: Chemical Mass Balance; PCA: Principal Components Analysis; FA: Factor Analysis; APCFA: Absolute Principal Component Factor Analysis; PMF: Positive Matrix Factorization; CPF: Conditional Probability Function; PSCF: Potential Source Contribution Function; ME: Multilinear Engine; COPREM: Constrained Physical Receptor Model.

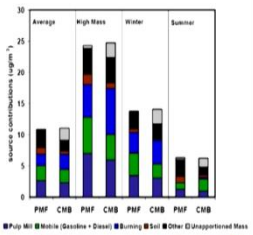
Sources: (Belis et al., 2014)

Selection of the Receptor Models (2/3)

The most commonly used Receptor Models are Chemical Mass Balance (CMB) and Positive Matrix Factorization (PMF) models.

The CMB model is a 'least squares' model which estimates source contributions on the basis of the chemical fingerprints of the source and the concentration of pollutants.

The PMF model is based on uncertainty-weighted factor analysis which relies on pollutant measurements.



Sources: (Belis et al., 2014)

Image: (Barn et al., 2011)

Selection of the Receptor Models (3/3)

- The main strength of the CMB model is that, unlike other statistical receptor models (e.g., PMF), it does not require a large dataset and theoretical equation can be solved for an individual sample.
- The CMB output does not require additional identification of the contributing sources/factors, since the profiles are pre-selected for specific sources.

From Eq. 1

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} \quad \text{.....Eq. 2}$$

- Assume there is no residual term.

Sources: (Belis et al., 2014)

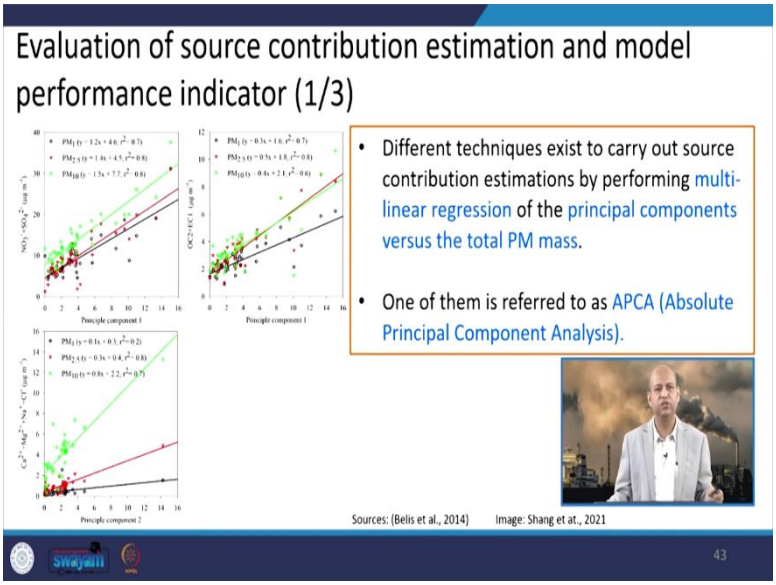
Well, when we want to select the receptor models, so there are many options like chemical mass balance, positive matrix factorization, factor analysis without constraints then hybrid expanded models. So, a number of models are there but we need to do which is more popular more important. And why do we select this CMB because this is the most commonly used receptor model, this chemical mass balance and positive matrix factorization PMF.

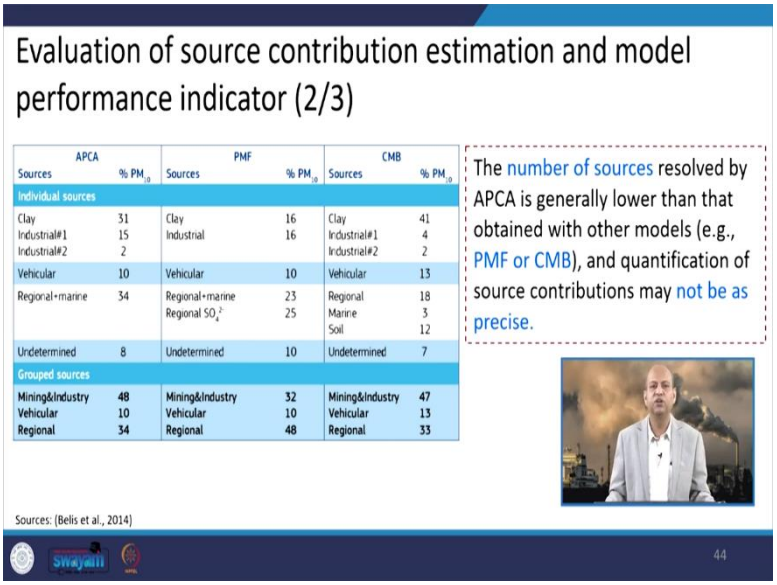
CMB and PMF are most popular models they are easy to use and CMB model is the least square model basically which estimates the source contribution on the basis of the chemical fingerprints of the source and the concentration of the pollutants. So, that relationship is easy to establish and this positive matrix factorization model is based on uncertainty weighted factor analysis which relies on pollutant measurement.

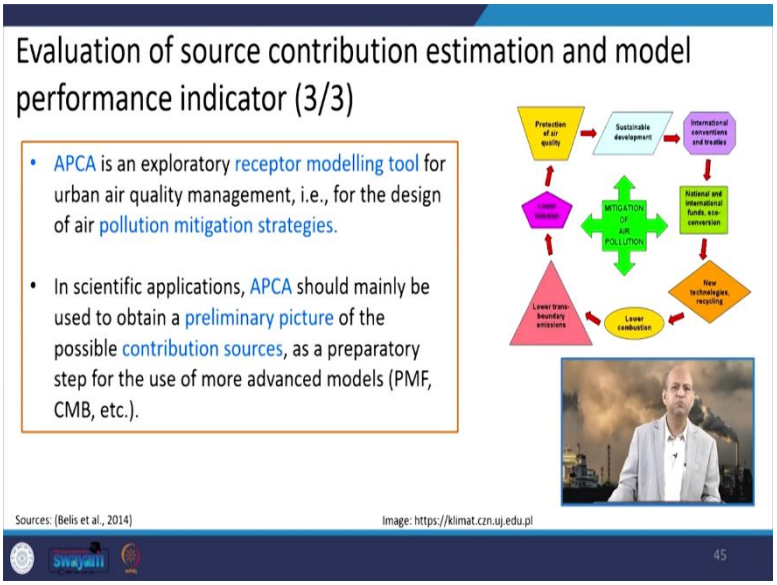
So, this is the basic difference between two and the main strength of the CMB model is that unlike other statistical receptor models like PMF, if you compare CMB and PMF, then CMB is better than PMF because it does not require a large data set. Wherever in PMF you need to have large data set and the theoretical equation can be solved for an individual sample.

That is the beauty of the CMB and the CMB output does not require additional identification of the contributing sources factors since the profiles are pre-selected and for specific sources. So, with this equation we have already seen, so this is the basic equation for CMB analysis.

(Refer Slide Time: 36:24)





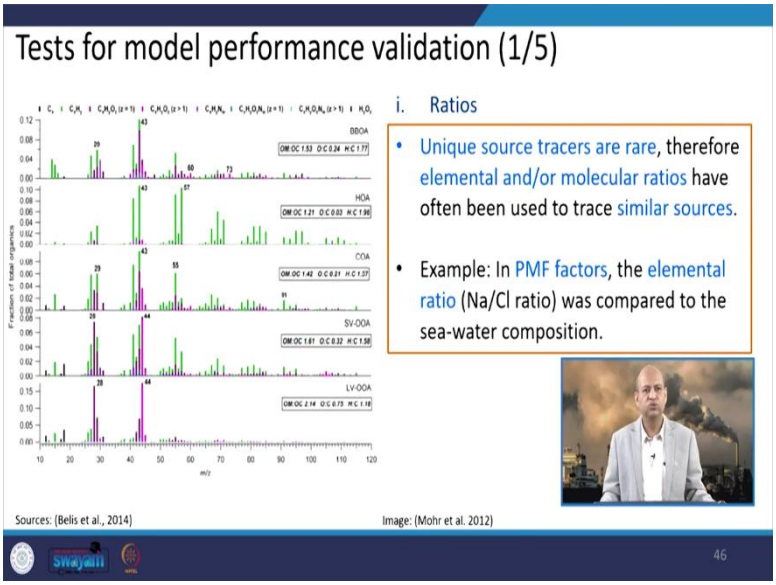


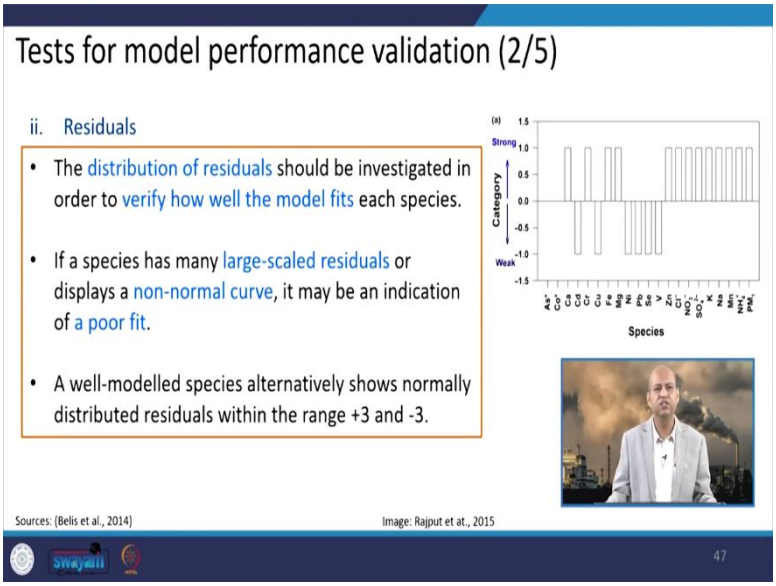
When we go for evaluation of source contribution estimation and model performance indicators, then there are various indicators but one of the important indicator is Absolute Principle Component Analysis, (APCA). So, you can see this APCA values for different like CMB, PMF are shown and the number of sources which are resolved by APCA is generally lower than the obtained with other models like PMF or CMB.

And quantification of source contribution may not be as precise. So, that is one particular feature of this and this APCA is an exploratory receptor modeling tool for urban air quality management, for the design of air pollution mitigation strategies. So, this is very popular in that sense and the scientific application is, in scientific application APCA should mainly be used to obtain preliminary picture of the possible contribution sources.

So, for detailed one it is not so good but for preliminary it is good, as a preparatory step for use of more advanced model PMF, CMB. So, before that APCA can be used for preliminary analysis.

(Refer Slide Time: 37:41)

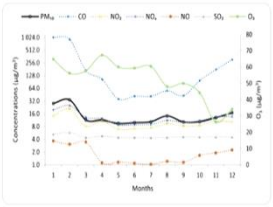




Tests for model performance validation (3/5)

iii. Time trends

- Source strengths are often time-dependent due to the influence of atmospheric processes (nucleation, volatilisation, transport, etc.), meteorological parameters (solar radiation, humidity, precipitation, etc.), and variation in human activity (intra-day, day-to-day).
- As a result, the source contributions will also change over time, and this variation is a suitable diagnostic for evaluating interpretations of factor profiles.



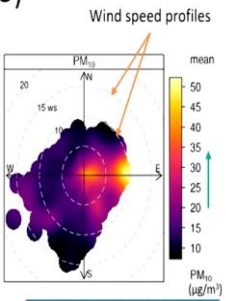
Sources: (Belis et al., 2014)

Image: Bodor et al., 2020

Tests for model performance validation (4/5)

iv. Wind direction analysis

- A simple but reliable method is to plot source contributions in a polar scatter plot in such a way that wind direction determines the angle and source contribution determines the radius of each plotted point.
- Such a plot shows briefly the general behavior of wind-directional dependence.



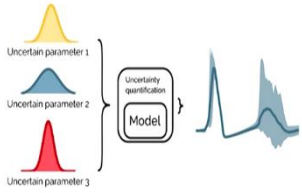
Sources: (Belis et al., 2014)

Image: (Grange, Lewis, and Carslaw 2016)

Tests for model performance validation (5/5)

v. Overall uncertainty

- The output from source apportionment (SA) consists of source contribution estimates (SCEs) with a definite uncertainty.
- Most receptor models compute the uncertainty of the output.
- Where results derive from more than one SA technique, the computation of the uncertainty of the combined SCEs is not easy.



Sources: (Belis et al., 2014)

Image: https://www.frontiersin.org

So, the test of these model for performance this validation, you can do various techniques you can use like ratios unique source tracers are there which are rare and therefore element and molecular ratios have often been used for this trace similar sources and examples are like PMF factors, the elemental ratio like sodium and chloride ratio and compared to the sea water composition. So, that relationship can be easily established.

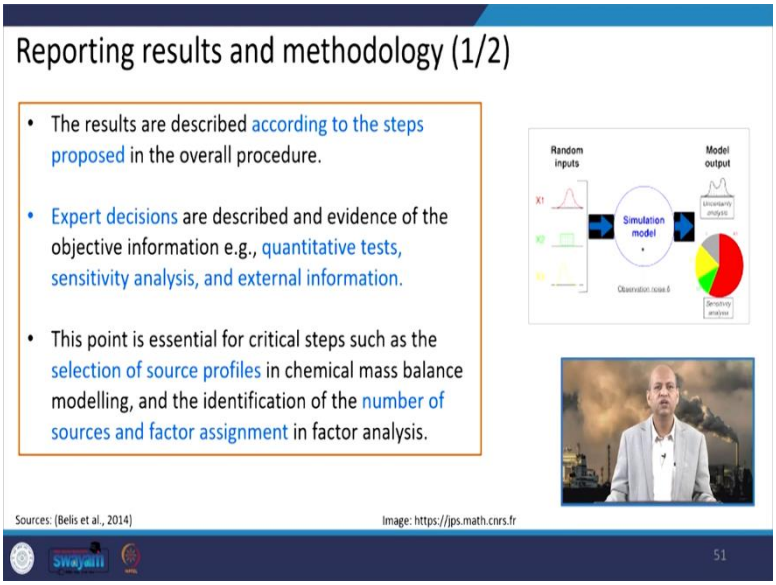
There may be some residuals for normal curve, that is also one important. And the well-modeled species alternatively shows normally distributed residual within the range of + 3 and - 3. So, that is also important because the distribution of residuals should be investigated in order to verify how well the model is fitting with each species. So, that is also important.

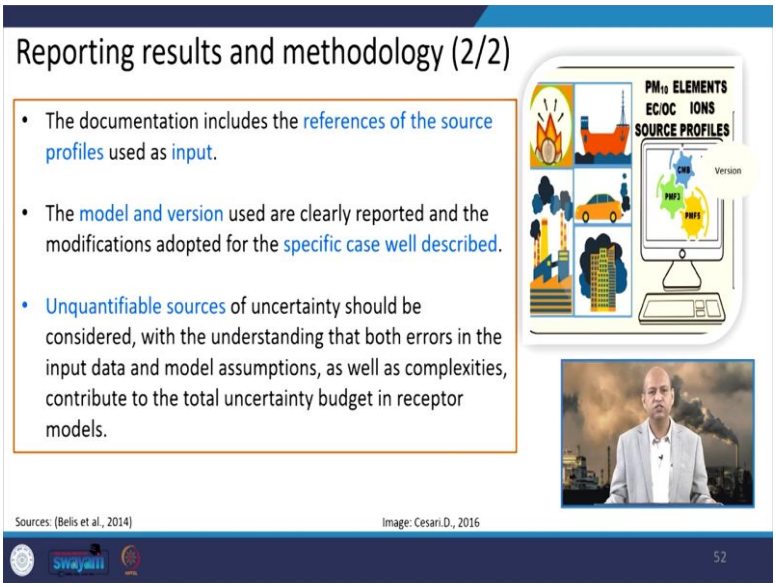
Time trends which we have already discussed but again we can see that time dependent due to the influence of atmospheric processes and meteorological parameters variation due to some human activity also. Time trends may be there and we can know why the variation is there, what is the reason is there. So, as a result the source contribution will also change over time like variation is suitable diagnostic for evaluating interpretation of factor profiles according to the time relation.

When we talk about wind direction analysis, so this is the plot where wind profile is also there, pollution profile is also there. You might be knowing this wind rose diagram so those wind rose diagram and pollution rose diagram are clubbed here and this is a kind of simple but reliable method to plot source contributions in a polar scatter plot which is shown here. And it is plotted in such a way that the wind direction determines the angle and the source contribution determines the radius of each of the plot.

And such a plot shows briefly the general behaviour of wind directional dependence. So, maybe from a particular wind direction more pollution is coming so that insight is very much available from these kind of analysis. When we talk about overall uncertainty then again we have to do some source of uncertainty analysis, so that we can figure out, what is the reliability of those data which we are getting out of this analysis.

(Refer Slide Time: 40:10)





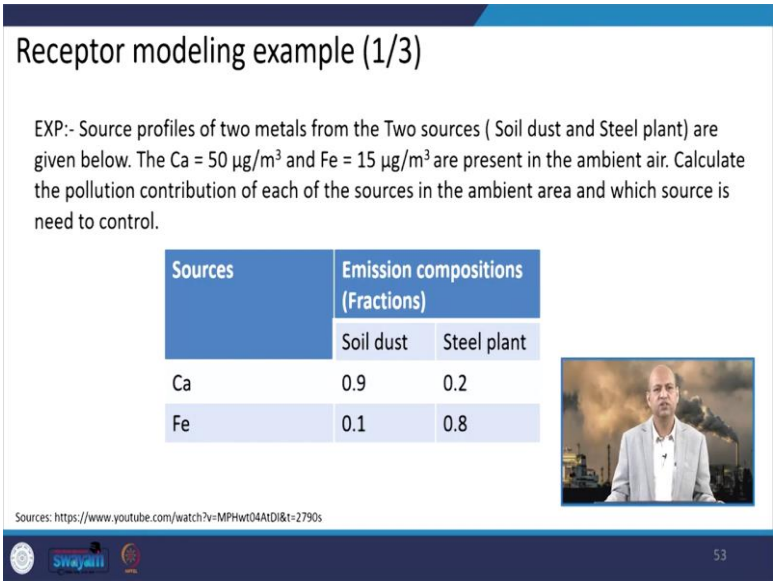
And the results which are described according to the steps proposed for overall procedures are to be reported in a nice manner. So, when we talk about reporting the results and methodology it is also important how do you report those results which you have got with this whole extensive analysis.

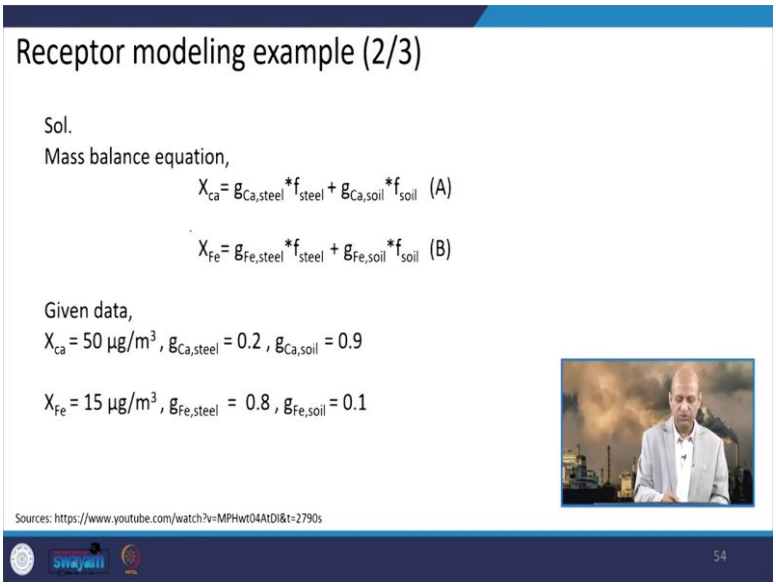
So, experts decision are described and evidence of the objective information like quantitative test or sensitivity analysis or uncertainty analysis, external information all these things should be taken into account and incorporated in the report and presentation of those results. So, this point is essential for critical steps such as selection of source profile in chemical mass balance modeling and the identification of the number of sources and factors assigned in factor analysis. So, that is why this is important.

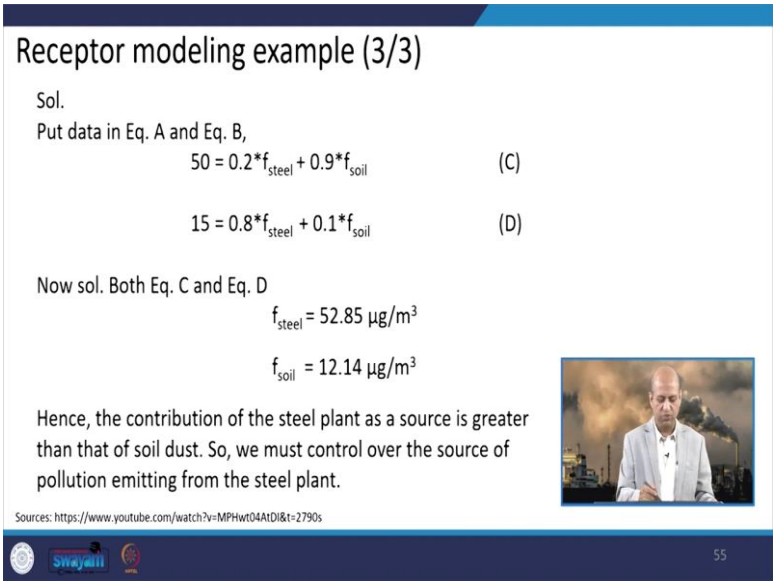
Then when we do documentation so it should include like references of the source profiles and the input and the model and their version which model you are using which latest version or which version you are using for the analysis because that will also influence the results preciseness and the specific case which you are describing and then when we talk about unquantifiable sources of uncertainty, that uncertainty should be considered which are not quantifiable because they are there in actual sense but we are not considering and they will include some uncertainty in our results.

So, with the understanding of these errors in the input data and model assumptions as well as complexities, they contribute to the total uncertainty budget in a receptor model. So, that is also to be incorporated and properly discussed.

(Refer Slide Time: 41:56)





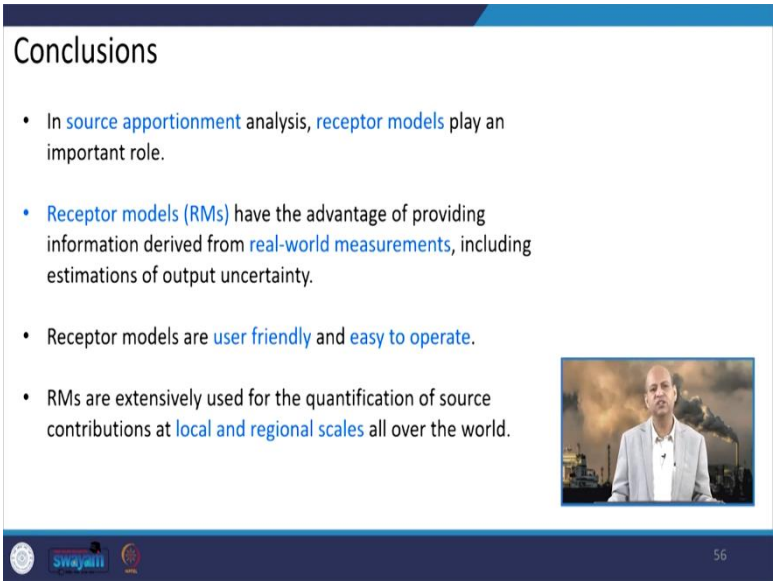


When we talk about sources receptor modeling, one example so these sources like soil dust or steel plant is there, so these sources, they are giving like calcium 50 microgram per cubic meter, iron fe 15 microgram per cubic meter. So, they are present in the ambient air. So, we need to calculate the pollution contribution of each of the sources in the ambient area and which sources need to be controlled, if we want to know.

And these are the fractions emission composition, 0.9, 0.2, 0.1, and 0.8. So, we use by that particular equation, so this mass balance equation we are using. This A and B for first calcium for iron and then put those values which are available in the problem and you calculate so you get these values and you can say like fraction in the steel 52.8 from the soil it is 12.14.

So, the contribution of the steel plant as a source is greater than that of the soil dust. So, we must control over the source of the pollution emitting from the steel plant. So, this kind of information we receive from the these simple calculations which we are using.

(Refer Slide Time: 43:14)



So, in conclusion we can say that the source apportionment analysis receptor models play a very important role in carrying out the source apportionment analysis. And the receptor models have the advantages of providing information derived from real world measurements including like estimations of output uncertainty and these receptor models are user friendly, very easy to operate, they are easy to handle and receptor models are extensively used for the quantification of source contribution at local and regional scales all over the world.

So, they are very popular technique. So, I hope you know with this presentation, you know a bit about what is the importance of receptor modeling and how do we use it for carrying out the source apportionment.

(Refer Slide Time: 43:58)

Well these are the references for additional information. Thank you for your kind attention. See, you in the next lecture, see you again. Thanks.